Gesture semantics reconstruction based on motion capturing and complex event processing

By: Thies Pfeiffer, Florian Hofmann, Florian Hahn, Hannes Reiser, and Insa Röpke (Lawler)

Pfeiffer, T., Hofmann, F., Hahn, F., Rieser, H., Röpke, I. (2013). Gesture semantics reconstruction based on motion capturing and complex event processing. *14th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 270-279.
https://www.aclweb.org/anthology/W13-4041

Abstract:

A fundamental problem in manual based gesture semantics reconstruction is the specification of preferred semantic concepts for gesture trajectories. This issue is complicated by problems human raters have annotating fast-paced three dimensional trajectories. Based on a detailed example of a gesticulated circular trajectory, we present a data-driven approach that covers parts of the semantic reconstruction by making use of motion capturing (mocap) technology. In our FA^{3}ME framework we use a complex event processing approach to analyse and annotate multi-modal events. This framework provides grounds for a detailed description of how to get at the semantic concept of circularity observed in the data.

Keywords: semantics | complex event processing | gesture annotation | circular trajectory

Article:

*****Note: Full text of article below**

Gesture Semantics Reconstruction Based on Motion Capturing and Complex Event Processing: a Circular Shape Example

Thies Pfeiffer, Florian Hofmann
Artificial Intelligence Group
Faculty of Technology
Bielefeld University, Germany
(tpfeiffe|fhofmann)
@techfak.uni-bielefeld.de

Florian Hahn, Hannes Rieser, Insa Röpke
Collaborative Research Center
"Alignment in Communication" (CRC 673)
Bielefeld University, Germany
(fhahn2|hannes.rieser|iroepke)
@uni-bielefeld.de

Abstract

A fundamental problem in manual based gesture semantics reconstruction is the specification of preferred semantic concepts for gesture trajectories. This issue is complicated by problems human raters have annotating fast-paced three dimensional trajectories. Based on a detailed example of a gesticulated circular trajectory, we present a data-driven approach that covers parts of the semantic reconstruction by making use of motion capturing (mocap) technology. In our FA^{3}ME framework we use a complex event processing approach to analyse and annotate multi-modal events. This framework provides grounds for a detailed description of how to get at the semantic concept of circularity observed in the data.

1 Introduction

Focussing on iconic gestures, we discuss the benefit of motion capturing (mocap) technology for the reconstruction of gesture meaning and speech meaning: A fundamental problem is the specification of semantic concepts for gesture trajectories, e.g., for describing circular movements or shapes. We start with demonstrating the limitations of our manual based annotation. Then we discuss two strategies of how to deal with these, pragmatic inference *vs.* low level annotation based on mocap technology yielding a more precise semantics. We then argue that the second strategy is to be preferred to the inferential one.

The annotation of mocap data can be realised semi-automatically by our FA^{3}ME framework for the analysis and annotation of multi-modal events, which we use to record multi-modal corpora. For mocap we use the tracking system ART DTrack2 (advanced realtime tracking GmbH, 2013), but the framework is not restricted to this technical set-up. In cooperation with others (e.g., (Kousidis et al., 2012)), we also have used products from Vicon Motion Systems (2013) and the Microsoft Kinect (Microsoft, 2013). Pfeiffer (2013) presents an overview on mocap technology for documenting multi-modal studies.

We thus provide details about the way gestures are analysed with FA^{3}ME and about the procedure to reconstruct the gesture meaning for the circular movement in our chosen example. We conclude with a discussion of how these low-level reconstructions can be integrated into the reconstruction of speech and gesture meaning.

2 From Linguistic Annotation to MoCap

In this section we describe our methodology for the reconstruction of gesture meaning, speech meaning and its interfacing, illustrated by an example. We then show a shortcoming of our corpus-based annotation and discuss two possible solutions to amend it, pragmatic inference *vs.* semantics based on mocap technology. The technology described in Section 3 will in the end enable us to get the preferred reconstruction of gesture semantics.

The reconstruction of the gesture meaning and its fusion with speech meaning to get a multimodal proposition works as follows: On the speech side we start with a manual transcription, upon which we craft a context free syntax analysis followed by a formal semantics. On the gesture side we build an AVM-based representation of the gesture resting on manual annotation.[1] Taking the gesture as a sign with independent meaning (Rieser, 2010), this representation provides the basis for the formal gesture semantics. In the next

[1] We do not use an explicit gesture model, which would go against our descriptive intentions. The range of admissible gestures is fixed by annotation manuals and investigations in gesture typology.

Proceedings of the SIGDIAL 2013 Conference, pages 270–279,
Metz, France, 22-24 August 2013.

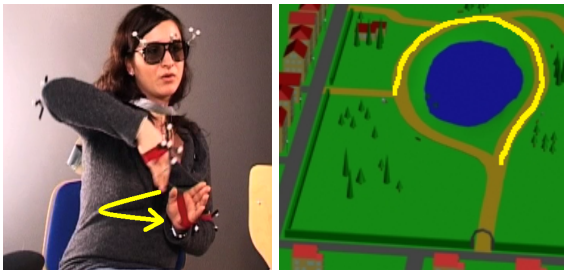

Figure 1: Our example: a circular gesture (left: video still) to describe the path around the pond (right).

step, the gesture meaning and the speech meaning are fused into an interface (Röpke et al., 2013). Every step in these procedures is infested by underspecification which, however, we do not deal with here. These are, for instance, the selection of annotation predicates, the attribution of logical form to gestures and the speech analysis.

In our example, we focus on two gesture parameters, the movement feature of the gesture and the representation technique used. It originates from the systematically annotated corpus, called SaGA, the Bielefeld Speech-and-Gesture Alignment-corpus (Lücking et al., 2012). It consists of 25 dialogues of dyads conversing about a "bus ride" through a Virtual Reality town. One participant of each dyad, the so-called Route-Giver (RG), has done this ride and describes the route and the sights passed to the other participant, the so-called Follower (FO). The taped conversations are annotated in a fine-grained way.

In the example, the RG describes a route section around a pond to the FO. While uttering "Du gehst drei Viertel rum/You walk three quarters around", she produces the gesture depicted in Figure 1. Intuitively, the gesture conveys a circularity information not expressed in the verbal meaning. In order to explicate the relation of speech and gesture meaning, we use our methodology as described above. To anticipate, we get a clear contribution of the speech meaning which is restricted by the gesture meaning conveying the roundness information. The first step is to provide a syntactical analysis which you can see in Figure 2.[2]

[2]The gesture stroke extends over the whole utterance. Verb phrases can feature so-called "sentence brackets". Here, due to a sentence bracket, the finite verb stem "gehst" is separated from its prefix ("rum"). Together they embrace the German *Mittelfeld*, here "drei Viertel". Observe the N-ellipsis "∅" in the NP. The prefix and the finite verb stem cannot be fully interpreted on their own and are therefore marked with an asterisk.

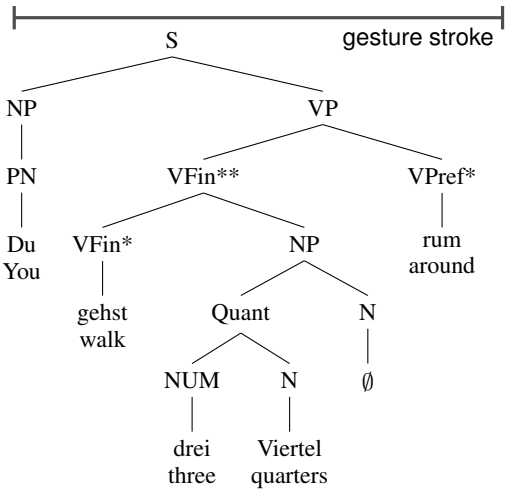

Figure 2: Syntax analysis

The speech representation is inspired by a Montague-Parsons-Reichenbach event ontology, and uses type-logic notions. Ignoring the embedding in an indirect speech act[3], the speech semantics represents an AGENT (the FO) who is engaged in a WALK-AROUND event e related to some path F, and a THEME relating the WALK-AROUND event e with the path F.

$$\exists e y \mathrm{F}\ 3/4x(\textsc{walk-around}(e) \wedge \textsc{agent}(e, \mathrm{FO}) \wedge \textsc{theme}(e, x) \wedge \mathrm{F}(x, y)) \quad (1)$$

The gesture semantics is obtained using the annotated gesture features. The relevant features are the movement of the wrist (Path_of_Wrist) and the Representation_Technique used.

$$\begin{bmatrix} \text{Path\_of\_Wrist} & \text{ARC<ARC<} \\ & \text{ARC<ARC} \\ \text{Representation\_Technique} & \text{Drawing} \end{bmatrix}^{4}$$

Interpreting the values ARC<ARC< ARC<ARC and Drawing, respectively, the calculated gesture semantics represents a bent trajectory consisting of four segments:

[3]We have treated the function of speech-gesture ensembles in dialogue acts and dialogues elsewhere (Rieser and Poesio (2009), Rieser (2011), Hahn and Rieser (2011), Lüecking et al. (2012)).

[4]This is a shortened version of the full gesture-AVM. Features like hand shape etc. are ignored. See Rieser (2010) for other annotation predicates.

$$\exists x y_1 y_2 y_3 y_4 (\text{TRAJECTORY}_0(x) \wedge \text{BENT}(y_1) \wedge \text{BENT}(y_2) \wedge \text{BENT}(y_3) \wedge \text{BENT}(y_4) \wedge y_1 < y_2 < y_3 < y_4 \wedge \text{SEGMENT}(y_1, x) \wedge \text{SEGMENT}(y_2, x) \wedge \text{SEGMENT}(y_3, x) \wedge \text{SEGMENT}(y_4, x)). \quad (2)$$

The paraphrase is: There exists a TRAJECTORY_0 x which consists of four BENT SEGMENTS y_1, y_2, y_3, y_4. We abbreviate this formula to:

$$\exists x_1 (\text{TRAJECTORY}_1(x_1)) \quad (3)$$

In more mundane verbiage: There is a particular TRAJECTORY_1 x_1. In a speech-gesture interface[5] (Rieser, 2010) both formulae are extended by adding a parameter in order to compositionally combine them:

$$\lambda Y. \exists e y \text{F } 3/4 x \, (\text{WALK-AROUND}(e) \wedge \text{AGENT}(e, \text{FO}) \wedge \text{THEME}(e, x) \wedge \text{F}(x, y) \wedge \text{Y}(y)) \quad (4)$$

We read this as: There is a WALK-AROUND event e the AGENT of which is FO related to a three quarters (path) F. This maps x onto y which is in turn equipped with property Y.

$$\lambda z. \exists x_1 (\text{TRAJECTORY}_1(x_1) \wedge x_1 = z) \quad (5)$$

This means "There is a TRAJECTORY_1 x_1 identical with an arbitrary z". The extensions (4) and (5) are based on the intuition that the preferred reading is a modification of the (path) F by the gesture.

Taking the gesture representation as an argument for the speech representation, we finally get a simplified multi-modal interface formula. The resulting proposition represents an AGENT (FO) who is engaged in a WALK-AROUND event e and a THEME that now is specified as being related to a bent trajectory of four arcs due to formula (2):

$$\exists e y \, 3/4 x \, \exists \text{F}(\text{WALK-AROUND}(e) \wedge \text{AGENT}(e, \text{FO}) \wedge \text{THEME}(e, x) \wedge \text{F}(x, y) \wedge \text{TRAJECTORY}_1(y)) \quad (6)$$

We take this to mean "There is an AGENT FO's WALK-AROUND event e related to a three quarters (path) F having a TRAJECTORY_1 y".

As a result, the set of models in which the orginal speech proposition is true is restricted to the set of models that contain a bent trajectory standing in relation to the (path) F. But this restriction is too weak. Intuitively, the gesture conveys the meaning of a horizontal circular trajectory and not just four bent arcs. To see the shortcoming, note that the set of models also includes models which include a path having four bends that do not form a circular trajectory.

We envisage two methods to get the appropriate circularity intuition: pragmatic enrichment and an improvement of our gesture datum to capture the additional information conveyed in the gesture: By pragmatic enrichment, on the one hand, horizontal orientation and circularity of the gesture trajectory are inferred using abduction or defaults. However, the drawback of working with defaults or abduction rules is that we would have to set up too many of them depending on the various shapes and functions of bent trajectories.

On the other hand, the datum can be improved to yield a circularly shaped trajectory instead of the weaker one consisting of four bent arcs. Our motion capture data supports the second method: The motion capture data allows us to compute the complete trajectory drawn in the gesture space. This will be the basis for producing a mapping from gesture parameters to qualitative relations which we need in the truth conditions. In the end, we achieve a circular trajectory that is defined as one approximating a circle, see Section 4.3.

In this mapping procedure resides an underspecification, which is treated by fixing a threshold for the application of qualitative predicates through raters' decisions. This threshold value will be used in giving truth conditions for, e.g., (11), especially for determining APPROXIMATE.

We prefer the second method since it captures our hypothesis that the gesture as a sign conveys the meaning *circular trajectory*. The gain of the automated annotation *via* mocap which we will see subsequently is an improvement of our original gesture datum to a more empirically founded one. As a consequence, the set of models that satisfy our multi-modal proposition can be specified. This is also the reason for explicitly focussing on gesture semantics in this paper.

3 FA^{3}ME - Automatic Annotation as Complex Event Processing

The creation of FA^{3}ME, our *Framework for the Automatic Annotation and Augmentation of Multi-*

[5] How our model deals with interfacing speech meaning and gesture meaning has been elaborated in a series of papers (see footnote 3). We are well aware of the work on gesture-speech integration by Lascarides and colleagues which we deal with in a paper on interfaces (Rieser (2013)).

modal Events, is *inter alia* motivated by our key insight from previous studies that human raters have extreme difficulties when annotating 3D gesture poses and trajectories. This is especially true when they only have a restricted view on the recorded gestures. A typical example is the restriction to a fixed number of different camera angles from which the gestures have been recorded. In previous work (Pfeiffer, 2011), we proposed a solution for the restricted camera perspectives based on mocap data: Our Interactive Augmented Data Explorer (IADE) allowed human raters to immerse into the recorded data via virtual reality technology. Using a 3D projection in a CAVE (Cruz-Neira et al., 1992), the raters were enabled to move freely around and through the recorded mocap data, including a 3D reconstruction of the experimental setting. This interactive 3D visualisation supported an advanced annotation process and improved the quality of the annotations but at high costs. Since then, we only know of Kipp (2010) who makes mocap data visible for annotators by presenting feature graphs in his annotation tool Anvil in a desktop-based setting. In later work, Nguyen and Kipp (2010) also support a 3D model of the speaker, but this needed to be handcrafted by human annotators. A more holistic approach for gesture visualizations are the Gesture Space Volumes Pfeiffer (2011), which summarize gesture trajectories over longer timespans or multiple speakers.

The IADE system also allowed us to add visual augmentations during the playback of the recorded data. These augmentations were based on the events from the mocap data, but aggregated several events to higher-level representations. In a study on pointing gestures (Lücking et al., 2013), we could test different hypotheses about the construction of the direction of pointing by adding visual pointing rays shooting in a 3D reconstruction of the original real world setting. This allowed us to asses the accuracy of pointing at a very high level in a data-driven manner and derive a new model for the direction of pointing (Pfeiffer, 2011).

3.1 Principles in FA^{3}ME

In the FA^{3}ME project, we iteratively refine our methods for analysing multi-modal events. As a central concept, FA^{3}ME considers any recorded datum as a *first-level multi-modal event* (see Figure 3, left). This can be a time-stamped frame from a video camera, an audio sample, 6-degree-of-freedom matrices from a mocap system or gaze information from an eye-tracking system (e.g., see Kousidis et al. (2012)).

A distinctive factor of FA^{3}ME is that we consider annotations as *second-level multi-modal events*. That is, recorded and annotated data share the same representation. Annotations can be added by both, human raters and classification algorithms (the event rules in Figure 3, middle). Annotations can themselves be target of annotations. This allows us, for example, to create automatic classifiers that rely on recorded data and manual annotations (e.g., the first yellow event in Figure 3 depends on first-level events above and the blue second-level event to the right). This is helpful when classifiers for complex events are not (yet) available. If, for instance, no automatic classifiers for the stroke of a gesture exists, these annotations can be made by human raters. Once this is done, the automatic classifiers can describe the movements during the meaningful phases by analysing the trajectories of the mocap data.

Third-level multi-modal events are augmentations or extrapolations of the data. They might represent hypotheses, such as in the example of different pointing rays given above.

3.2 Complex Event Processing

In FA^{3}ME, we consider the analysis of multi-modal events as a *complex event processing* (CEP) problem. CEP is an area of computer science dedicated to the timely detection, analysis, aggregation and processing of events (Luckham, 2002). In the past years, CEP has gained an increased attention especially in the analysis of business relevant processes where large amount of data, e.g., share prices, with high update rates are analysed. This has fostered many interesting tools and frameworks for the analysis of structured events (Arasu et al., 2004a; EsperTech, 2013; Gedik et al., 2008; StreamBase, 2013). Hirte et al. (2012) apply CEP to a motion tracking stream from a Microsoft Kinect for real-time interaction, but we know of no uses of CEP for the processing of multi-modal event streams for linguistic analysis.

Dedicated query languages have been developed by several CEP frameworks which allow us to specify our event aggregations descriptively at a high level of abstraction (Arasu et al., 2004b;

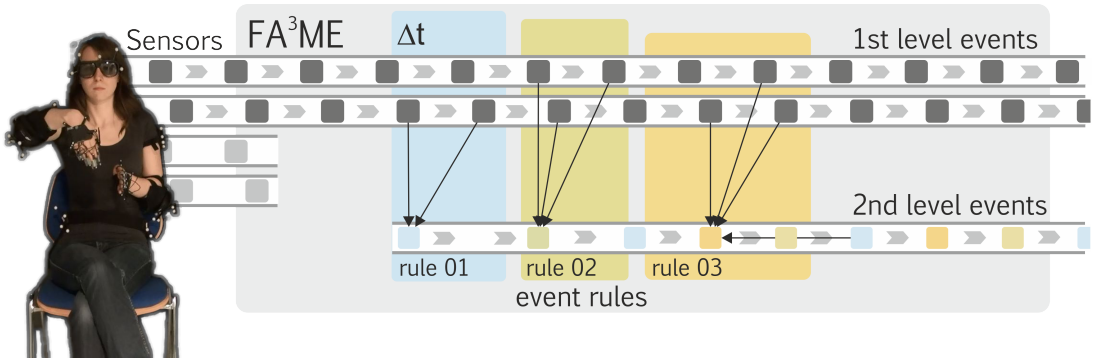

Figure 3: In FA^{3}ME, incoming multi-modal events are handled by a complex event processing framework that matches and aggregates events based on time windows to compose 2nd level multi-modal events. All multi-modal events can then be mapped to tiers in an annotation tool.

Gedik et al., 2008). The framework we use for FA^{3}ME is Esper (EsperTech, 2013), which provides a SQL-like query language. As a central extension of SQL, CEP query languages introduce the concept of event streams and time windows as a basis for aggregation (see Figure 3).

The CEP approach of FA^{3}ME allows us to create second- and third-level multi-modal events on-the-fly. We can thus provide near real-time annotations of sensor events. However, we have to consider the latencies introduced by sensors or computations and back-date events accordingly.

As a practical result, once we have specified our annotation descriptions formally in the language of CEP, these descriptions can be used to create classifiers that operate both on pre-recorded multi-modal corpora and on real-time data. This makes CEP interesting for projects where research in Linguistics and Human-Computer Interaction meet.

4 From MoCap to Linguistic Models

In this section, we will now address the problem of annotating the circular trajectory. In order to get the preferred semantics we yet cannot rely exclusively on the automatic annotation. We need the qualitative predicate "phase" to identify the meaningful part of the gesture (the stroke). Additionally, the qualitative predicate "representation technique" is required to select the relevant mocap trackers. For instance, the representation technique "drawing" selects the marker of the tip of the index finger. We thus require a hybrid model of manual and automatic annotations. In the following, we will focus on the automatic annotation.

First of all, when using mocap to record data, a frame of reference has to be specified as a basis for all coordinate systems. We chose a person-centered frame of reference anchored in the solar plexus (see Figure 4). The coronal plane is defined by the solar plexus and the two shoulders. The transverse plane is also defined by the solar plexus, perpendicular to the coronal plane with a normal-vector from solar plexus to the point ST (see Figure 4) between the two shoulders.

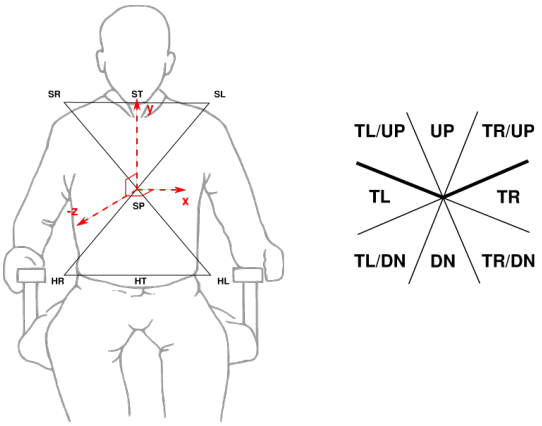

Figure 4: The coordinate system of the speaker (left). The orientations of the palms are classified into eight main directions (right).

4.1 Basic Automatic Gesture Annotations

The analysis of mocap data allows us to create basic annotations that we use in our corpora on-the-fly. This speeds up the annotation process and lets human raters focus on more complex aspects. One basic annotation that can be achieved automatically is the classification of the position of gesturing hands according to the gesture space model of McNeill (1992). As his annotation schema (see Figure 5, right) is tailored for the annotation of

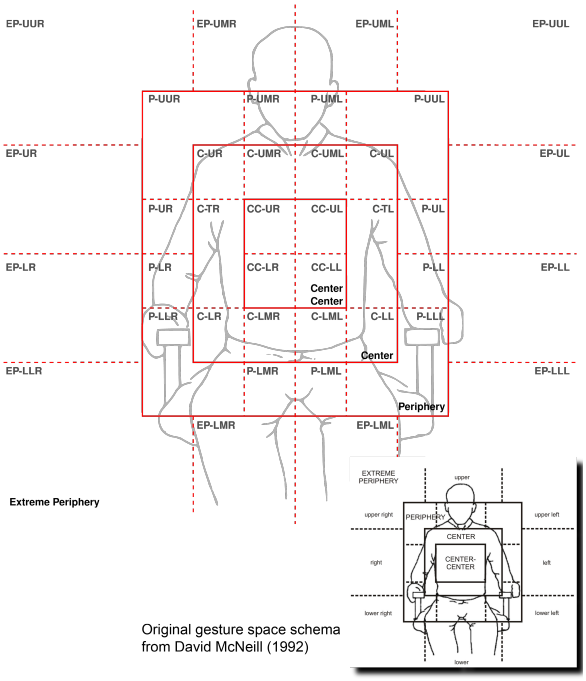

Figure 5: Our extended gesture space categorisation (upper left) is based on the work of McNeill (lower right).

video frames, we extended this model to support mocap as well (see Figure 5, left). The important point is that the areas of our schema are derived from certain markers attached to the observed participant. The upper right corner of the area C-UR (Center-Upper Right), for example, is linked to the marker for the right shoulder. Our schema thus scales directly with the size of the participant. Besides this, the sub-millimeter resolution of the mocap system also allows us to have a more detailed structure of the center area. The schema is also oriented according to the current coronal plane of the participant and not, e.g., according to the perspective of the camera.

A second example is the classification of the orientation of the hand, which is classified according to the scheme depicted in Figure 4, right. This classification is made relative to the transversal plane of the speaker's body.

4.2 Example: The Circular Trajectory

For the detection and classification of gestures drawing shapes two types of multi-modal events are required. First, multi-modal events generated by the mocap system for the hand. These events contain matrices describing the position and orientation of the back of the hand. Second, multi-modal events that mark the gesture stroke (one event for the start and one for the end) have to be generated, either by hand or automatically. At the moment, we rely on our manual annotations for the existing SaGA corpus.

We realise the annotation of circular trajectories in two steps. First, we reduce the trajectory provided by the mocap system to two dimensions. Second, we determine how closely the 2D trajectory approximates a circle.

Projection of the 3D Trajectory

The classifier for circles collects all events for the hand that happened between the two events for the stroke. As noted above, these events represent the position and orientation of the hand in 3D-space. There are several alternatives to reduce these three dimensions to two for classifying a circle (a 3D Object matching a 2D circle would be a sphere, a circular trajectory through all 3 dimensions a spiral). The principal approach is to reduce the dimensions by projecting the events on a 2D plane.

$$\exists xy\,(\textsc{trajectory}(x) \wedge \textsc{projection-of}(x, y) \wedge \textsc{trajectory2d}(y)) \quad (7)$$

Which plane to chose depends on the choice made for the annotation (e.g., global for the corpus) and thus on the context. For the description of gestures in dialogue there are several plausible alternatives. First, the movements could be projected onto one of the three body planes (sagittal plane, coronal plane, transversal plane). In our context, the transversal plane is suitable, as we are dealing with descriptions of routes, which in our corpus are made either with respect to the body of the speaker or with respect to the plane of an imaginary map, both extend parallel to the floor. Figure 6 (upper left) shows the circular movement in the transversal plane. A different perspective is presented in Figure 6 (right). There the perspective of a bystander is chosen. This kind of perspective can be useful for describing what the recipient of a dialogue act perceives, e.g., to explain misunderstandings. For this purpose, the gesture could also be annotated twice, once from the speaker's and once from the recipient's perspective.

At this point we want to emphasise that position and orientation of the planes do not have to be static. They can be linked to the reference points provided by the mocap system. Thus when the speaker turns her body, the sagittal, coronal and

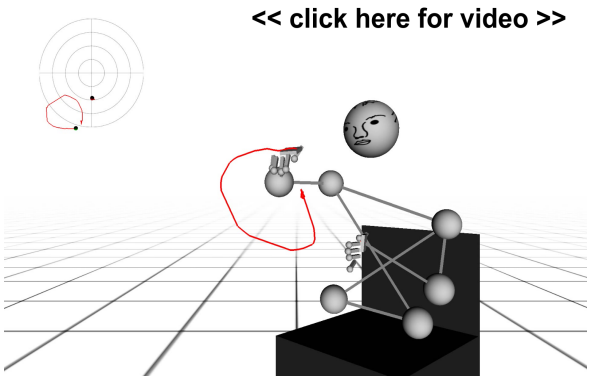

Figure 6: The circle-like gesture from our example can be visualised based on the mocap data. The right side shows the visualisation from the perspective of an interlocutor, the visualisation in the upper left corner is a projection of the movement on the transversal plane of the speaker.

transversal planes will move accordingly and the gestures are always interpreted according to the current orientation.

The plane used for projection can also be derived from the gesture itself. Using principal component analysis, the two main axes used by the gesture can be identified. These axes can then have arbitrary orientations. This could be a useful approach whenever 3D objects are described and the correct position and orientation of the ideal circle has to be derived from the gesture.

Circle Detection

Once the gesture trajectory has been projected onto a 2D plane, the resulting coordinates are classified. For this, several sketch-recognition algorithms have been proposed (e.g., (Alvarado and Davis, 2004; Rubine, 1991)). These algorithms have been designed for sketch-based interfaces (such as tablets or digitisers), either for recognising commands or for prettifying hand-drawn diagrams. However, once the 3D trajectory has been mapped to 2D, they can also be applied to natural gestures. The individual sketch-recognition algorithms differ in the way they are approaching the classification problem. Many algorithms follow a feature-based approach in which the primitives to be recognised are described by a set of features (such as aspect ratio or ratio of covered area) (Rubine, 1991). This approach is especially suited, when new primitives are to be learned by example. An alternative approach is the model-based approach in which the primitives to be recognised are described based on geometric models (Alvarado and Davis, 2004; Hammond and Davis, 2006). Some hybrid approaches also exist (Paulson et al., 2008). The model-based approaches are in line with our declarative approach to modelling, and are thus our preferred way for classifying shapes.

In our case, the projected 2D trajectory of the gesture is thus classified by a model-based sketch-recognition algorithm, which classifies the input into one of several shape classes (circle, rectangle, ...) with a corresponding member function $\text{ISSHAPE}(y, \text{CIRCLE}) \in [0 \ldots 1]$. By this, we can satisfy a subformula $\text{APPROXIMATES}(y, z) \wedge \text{CIRCLE}(z)$ by pre-setting a certain threshold. The threshold has to be chosen by the annotators, e.g., by rating positive and negative examples, as it may vary between participants and express the sloppiness of their gestures.

4.3 From MoCap to a Revision of Semantics

The result of the FA^{3}ME reconstruction of our gesture example can be expressed as follows:

$$\exists xyz\ (\text{TRAJECTORY}(x) \wedge \text{PROJECTION-OF}(x, y) \wedge \text{TRAJECTORY2D}(y) \wedge \text{APPROXIMATES}(y, z) \wedge \text{CIRCLE}(z)) \quad (8)$$

So we have: There is a projection of TRAJECTORY x, TRAJECTORY2D y, which is approximating a circle. We can now provide a description of the domain which can satisfy formula (8). Consequently, formula (8) is enhanced by definition (9).

$$\text{CIRCULAR\_TRAJECTORY}(x) =_{\text{DEF}} \exists yz(\text{TRAJECTORY}_2(x) \wedge \text{PROJECTION-OF}(x, y) \wedge \text{APPROXIMATES}(y, z) \wedge \text{circle}(z)) \quad (9)$$

This definition reads as "a CIRCULAR TRAJECTORY x is a TRAJECTORY$_2$ which has a PROJECTION y that approximates some circle z".

The formula (9) substitutes the TRAJECTORY$_1$ notion. The improved multi-modal meaning is (10):

$$\exists ey\, 3/4x\ \exists \text{F}(\text{WALK-AROUND}(e) \wedge \text{AGENT}(e, \text{FO}) \wedge \text{THEME}(e, x) \wedge \text{F}(x, y) \wedge \text{CIRCULAR\_TRAJECTORY}(y)) \quad (10)$$

Interfacing the new gesture representation with the speech representation captures our intuition that the gesture reduces the original set of models to a set including a circular-shaped trajectory.

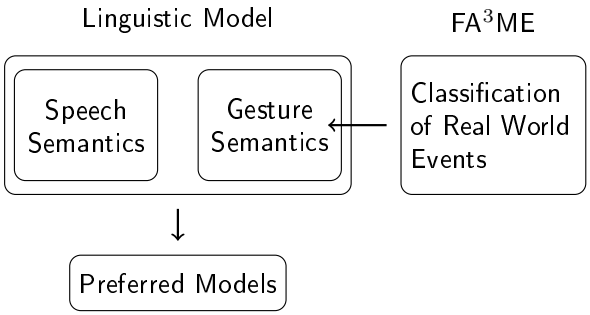

Figure 7: Specification of gesture semantics due to results of classification in FA^{3}ME. Simulation data feed into the gesture semantics which interfaces with the speech semantics.

The division of labour between linguistic semantics and FA^{3}ME technology regarding the semantic reconstruction is represented in Figure 7.

By way of explanation: We have the multi-modal semantics integrating speech semantics and gesture semantics accomplished *via* λ-calculus techniques as shown in Section 2. As also explained there, it alone would be too weak to delimit the models preferred with respect to the gesture indicating roundness. Therefore FA^{3}ME technology leading to a definition of CIRCULAR_TRAJECTORY is used which reduces the set of models to the preferred ones assuming a threshold n for the gestures closeness of fit to a circle. Thus, the relation between some gesture parameters and qualitative relations like *circular* can be considered as a mapping, producing values in the range $[0 \ldots 1]$. Still, it could happen that formula (8) cannot be satisfied in the preferred models. As a consequence, the multi-modal meaning would then fall short of satisfaction.

5 Conclusion

During our work on the interface between speech and gesture meaning our previous annotations turned out to be insufficient to support the semantics of concepts such as CIRCULAR_TRAJECTORY. This concept is a representative of many others that for human annotators are difficult to rate with the rigidity required for the symbolic level of semantics. Scientific visualisations, such as depicted in Figure 6, can be created to support the human raters. However, there is still the problem of perspective distortions three dimensional gestures are subject to when viewed from different angles and in particular when viewed on a 2D screen. It is also difficult to follow the complete trajectory of such gestures over time. Thus, one and the same gesture can be rated differently depending on the rater, while an algorithm with a defined threshold is not subject to these problems.

The presented hybrid approach based on qualitative human annotations, mocap and our FA^{3}ME framework is able to classify the particular 2D trajectories we are interested in following a three-step process: After the human annotator identified the phase and selected relevant trackers, the dimensions are reduced to two and a rigid model-based sketch-recognition algorithm is used to classify the trajectories. This classification is repeatable, consistent and independent of perspective. A first comparison of the manually annotated data and the automatic annotations revealed a high match. All differences between the annotations can be explained by restrictions of the video data which yielded a lower precision in the human annotations specifying the slant of the hand. Thus, the main issues we had with the results of human raters have been addressed, however a more formal evaluation on a large corpus remains to be done. What also remains is a specification of membership functions for each kind of gesture trajectories of interest (e.g., circular, rectangular, etc.). For this, a formal specification of what we commonly mean by, for instance, CIRCULAR, RECTANGULAR etc. is required.

The automated annotation *via* mocap improves our original gesture datum to capture the circularity-information conveyed in the gesture. We have a better understanding of the gesture meaning adopted *vis-à-vis* the datum considered. As it turns out, resorting to pragmatic inference cannot be avoided entirely, but we will exclude a lot of unwarranted readings which the manual-based logical formulae would *still* allow by using the approximation provided by body tracking methods. Not presented here is the way third-level multi-modal events are generated by re-simulating the data in a 3D world model to generate context events, e.g., to support pragmatics.

Acknowledgments

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673, *Alignment in Communication*. We are grateful to three reviewers whose arguments we took up in this version.

References

[advanced realtime tracking GmbH2013] A.R.T. advanced realtime tracking GmbH. 2013. Homepage. Retrieved May 2013 from http://www.ar-tracking.de.

[Alvarado and Davis2004] Christine Alvarado and Randall Davis. 2004. SketchREAD: a multi-domain sketch recognition engine. In *Proceedings of the 17th annual ACM symposium on User interface software and technology*, UIST '04, pages 23–32, New York, NY, USA. ACM.

[Arasu et al.2004a] Arvind Arasu, Brian Babcock, Shivnath Babu, John Cieslewicz, Mayur Datar, Keith Ito, Rajeev Motwani, Utkarsh Srivastava, and Jennifer Widom. 2004a. Stream: The stanford data stream management system. Technical report, Stanford InfoLab.

[Arasu et al.2004b] Arvind Arasu, Shivnath Babu, and Jennifer Widom. 2004b. CQL: A language for continuous queries over streams and relations. In *Database Programming Languages*, pages 1–19. Springer.

[Cruz-Neira et al.1992] Carolina Cruz-Neira, Daniel J. Sandin, Thomas A. DeFanti, Robert V. Kenyon, and John C. Hart. 1992. The cave: audio visual experience automatic virtual environment. *Communications fo the ACM 35 (2)*, 35(6):64–72.

[EsperTech2013] EsperTech. 2013. Homepage of Esper. Retrieved May 2013 from http://esper.codehaus.org/.

[Gedik et al.2008] Bugra Gedik, Henrique Andrade, Kun-Lung Wu, Philip S Yu, and Myungcheol Doo. 2008. SPADE: The System S Declarative Stream Processing Engine. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1123–1134. ACM.

[Hahn and Rieser2011] Florian Hahn and Hannes Rieser. 2011. Gestures supporting dialogue structure and interaction in the Bielefeld speech and gesture alignment corpus (SaGA). In *Proceedings of SEMdial 2011, Los Angelogue, 15th Workshop on the Semantics and Pragmatics of Dialogue*, pages 182–183, Los Angeles, California.

[Hammond and Davis2006] Tracy Hammond and Randall Davis. 2006. LADDER: A language to describe drawing, display, and editing in sketch recognition. In *ACM SIGGRAPH 2006 Courses*, page 27. ACM.

[Hirte et al.2012] Steffen Hirte, Andreas Seifert, Stephan Baumann, Daniel Klan, and Kai-Uwe Sattler. 2012. Data3 – a kinect interface for OLAP using complex event processing. *Data Engineering, International Conference on*, 0:1297–1300.

[Kipp2010] Michael Kipp. 2010. Multimedia annotation, querying and analysis in anvil. *Multimedia information extraction*, 19.

[Kousidis et al.2012] Spyridon Kousidis, Thies Pfeiffer, Zofia Malisz, Petra Wagner, and David Schlangen. 2012. Evaluating a minimally invasive laboratory architecture for recording multimodal conversational data. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog, INTERSPEECH2012 Satellite Workshop*, pages 39–42.

[Luckham2002] David Luckham. 2002. *The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems*. Addison-Wesley Professional.

[Lücking et al.2012] Andy Lücking, Kirsten Bergman, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2012. Data-based analysis of speech and gesture: the Bielefeld Speech and Gesture Alignment corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces*, -:1–14.

[Lücking et al.2013] Andy Lücking, Thies Pfeiffer, and Hannes Rieser. 2013. Pointing and reference reconsidered. *International Journal of Corpus Linguistics*. to appear.

[McNeill1992] David McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago.

[Microsoft2013] Microsoft. 2013. Homepage of KINECT for Windows. Retrieved May 2013 from http://www.microsoft.com/en-us/kinectforwindows/develop/.

[Nguyen and Kipp2010] Quan Nguyen and Michael Kipp. 2010. Annotation of human gesture using 3d skeleton controls. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*. ELDA.

[Paulson et al.2008] Brandon Paulson, Pankaj Rajan, Pedro Davalos, Ricardo Gutierrez-Osuna, and Tracy Hammond. 2008. What!?! no Rubine features?: using geometric-based features to produce normalized confidence values for sketch recognition. In *HCC Workshop: Sketch Tools for Diagramming*, pages 57–63.

[Pfeiffer2011] Thies Pfeiffer. 2011. *Understanding Multimodal Deixis with Gaze and Gesture in Conversational Interfaces*. Berichte aus der Informatik. Shaker Verlag, Aachen, Germany, December.

[Pfeiffer2013] Thies Pfeiffer. 2013. Documentation with motion capture. In Cornelia Müller, Alan Cienki, Ellen Fricke, Silva H. Ladewig, David McNeill, and Sedinha Teendorf, editors, *Body-Language-Communication: An International Handbook on Multimodality in Human Interaction*, Handbooks of Linguistics and Communication Science. Mouton de Gruyter, Berlin, New York. to appear.

[Rieser and Poesio2009] Hannes Rieser and M. Poesio. 2009. Interactive Gesture in Dialogue: a PTT Model. In P. Healey, R. Pieraccini, D. Byron, S. Yound, and M. Purver, editors, *Proceedings of the SIGDIAL 2009 Conference*, pages 87–96.

[Rieser2010] Hannes Rieser. 2010. On factoring out a gesture typology from the Bielefeld Speech-And-Gesture-Alignment corpus (SAGA). In Stefan Kopp and Ipke Wachsmuth, editors, *Proceedings of GW 2009: Gesture in Embodied Communication and Human-Computer Interaction*, pages 47–60, Berlin/Heidelberg. Springer.

[Rieser2011] Hannes Rieser. 2011. Gestures indicating dialogue structure. In *Proceedings of SEMdial 2011, Los Angelogue, 15th Workshop on the Semantics and Pragmatics of Dialogue*, pages 9–18, Los Angeles, California.

[Rieser2013] Hannes Rieser. 2013. Speech-gesture Interfaces. An Overview. In Heike Wiese and Malte Zimmermann, editors, *Proceedings of 35th Annual Conference of the German Linguistic Society (DGfS), March 12-15 2013 in Potsdam*, pages 282–283.

[Röpke et al.2013] Insa Röpke, Florian Hahn, and Hannes Rieser. 2013. Interface Constructions for Gestures Accompanying Verb Phrases. In Heike Wiese and Malte Zimmermann, editors, *Proceedings of 35th Annual Conference of the German Linguistic Society (DGfS), March 12-15 2013 in Potsdam*, pages 295–296.

[Rubine1991] Dean Rubine. 1991. Specifying gestures by example. In *Proceedings of the 18th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '91, pages 329–337, New York, NY, USA. ACM.

[StreamBase2013] StreamBase. 2013. Homepage of StreamBase. Retrieved May 2013 from `http://www.streambase.com/`.

[Vicon Motion Systems2013] Vicon Motion Systems. 2013. Homepage. Retrieved May 2013 from `http://www.vicon.com`.